Generalized Additive Partial Linear Models for Clustered Data with Diverging Number of Covariates Using GEE

Heng Lian, Hua Liang and Lan Wang

Nanyang Technological University, University of Rochester and University of Minnesota

Abstract: We study flexible modeling of clustered data using marginal generalized additive partial linear models with a diverging number of covariates. Generalized estimating equations are used to fit the model with the nonparametric functions being approximated by polynomial splines. We investigate the asymptotic properties in a "large n, diverging p" framework. More specifically, we establish the consistency and asymptotic normality of the estimators for the linear parameters under mild conditions. We propose a penalized estimating equations based procedure for simultaneous variable selection and estimation. The proposed variable selection procedure enjoys the oracle property and allows the number of parameters in the linear part to increase at the same order as the sample size under some general conditions. Extensive Monte Carlo simulations demonstrate that the proposed methods work well with moderate sample sizes. A dataset is analyzed to illustrate the application.

Key words and phrases: Clustered data, GEE, high dimension, injective function, marginal regression, Polynomial splines.

1. Introduction

Clustered data often arise in biological and biomedical research, where the measurements within the same cluster are correlated, while the measurements from different clustered are independent. For example, in longitudinal studies, the subjects are measured repeatedly over a given period of time. The measurements from the same subject are often correlated and thus form a cluster. A popular approach for clustered data analysis is generalized estimating equations (Liang and Zeger (1986)), in which both the within-cluster and between-cluster variations are considered. A remarkable property of the GEE estimator is that it is consistent and asymptotically normal even with a misspecified covariance matrix. Furthermore, the estimator is efficient when the covariance matrix is correctly specified. GEE and its extensions have been thoroughly studied for various parametric and semiparametric models and are broadly applied in diverse disciplines, see Diggle et al. (2002) for a comprehensive survey. Robust estimation based on

weighted GEE has been investigated in He, Fang and Zhu (2005).

Most existing work on GEE assumes the classical setting where the number of covariates is fixed. Clustered data involving high-dimensional covariates have become increasingly common from large-scale long-term health studies and from time-course gene expression experiments. For example, in the well known Framingham Heart Study, many covariates including age, smoking status, cholesterol level, blood pressure were recorded on the participants over the years to investigate their cardiovascular health. Sometimes, although the number of covariates are not many, when we consider various interaction effects, the total number of variables in the model is large. These modern applications motivate us to study analysis of clustered data in a new asymptotic setup, which allows the number of covariates to increase with the sample size. Wang (2011) recently studied the theory of GEE in a "large n, diverging p" asymptotic framework and revealed that most of the classical theory continues to hold under some general regularity conditions, see also Wang, Zhou and Qu (2012) for an extension to high-dimensional variable selection. However, the GEE model considered in Wang (2011) is restricted to a marginal linear model. Ma, Song and Wang (2012) considered additive partially linear longitudinal models. These are, however, limited to continuous response models and fixed dimensionality. To incorporate nonlinearity and avoid the curse of dimensionality, we study the marginal generalized additive partial linear models (GAPLM, Härdle et al. (2004); Wood (2006)) analysis of clustered data with diverging number of covariates.

The marginal GAPLM approach for clustered data analysis relaxes the restrictive model assumptions of marginal linear GEE. However, the more complex model structure, which involves both parametric and nonparametric components, also great computational challenges when the dimension of the covariates is high. More specifically, diverging dimension of linear components incorporating nonparametric modeling can be computationally intensive. In the setting of clustered data, this makes it difficult to incorporate additional correlation structure into the model. However, ignoring correlation leads to inefficient estimation and reduces prediction capability. As pointed out by Wang (2003), ignoring the correlation could also lead to biased estimation (Zhu, Fung and He (2008)). Specifically, Wang (2003) shows that selection of the smoothing parameter could fail since it is sensitive to even a small departure from the true correlation structure, and this is reflected by over-fitting the nonparametric estimator in order to reduce the overall bias. In contrast to the parametric setting, these problems could be

more serious for the GAPLM since the true model might not be easily verified. To the best of our knowledge, few efforts have been made in estimation of marginal generalized partially linear models (only one nonparametric function in GAPLM). For example, Lin and Carroll (2001, 2006) developed a GEE type estimating equation for such a setting, but no result for marginal GAPLM with diverging number of covariates is available.

For independent data, several algorithms have been proposed for estimation in GAPLM. The kernel-based backfitting or local scoring procedures (Buja, Hastie and Tibshirani (1989)) iteratively estimates the linear coefficients and nonparametric components by solving a large system of equations (Yu, Park and Mammen (2008)). The marginal integration approach (Linton and Nielsen (1995)) estimates the parametric components by treating the summand of additive terms as a nonparametric component, which is then estimated as a multivariate nonparametric function. Wood (2004) suggested penalized regression splines, which share most of the practical benefits of smoothing spline methods combined with ease of use and reduction of the computational cost of backfitting GAMs. Nevertheless, these methods all have limitations for estimating GAPLM when the dimension of covariates is large. The kernel-based backfitting procedure suffers from expensive computational costs because the procedure needs to solve a large system of equations (Yu et al. (2008)); while the marginal integration approach suffers from the "curse of dimensionality" (Härdle et al. (2004)). The difficulty increases dramatically as the dimension of covariates grows. The penalized spline estimators may not be efficient. Moreover, no theoretical justifications are available for these procedures even in the case the dimension of the covariates is fixed.

To fit GAPLM to clustered data with diverging number of covariates, we extend the spline-based approach proposed by Wang et al. (2011) for independent data. This approach uses polynomial splines to estimate the nonparametric components (Stone (1986, 1994)). Unknown functions are approximated via polynomial splines characterized by a linear combination of spline basis, and the coefficients of the linear part can be estimated by an efficient one-step procedure that maximizes the quasi-likelihood function after using the spline approximation to the nonparametric components. Such an approximation reduces the computational burden comparing to the local scoring backfitting approach or the marginal integration approach. To incorporate the intra-cluster correlation structure, we combine this algorithm with the GEE, which archives a good balance of incorporating the correlation structure and retaining numerical efficiency.

We establish the theory for marginal GAPLM analysis to clustered data in the "large n, diverging p" asymptotic framework. The theoretical development is quite challenging because of the curse of dimensionality of the nonparametric functions, the nonlinear relationship between the response and the covariates, and the intra-cluster correlation.

We consider variable selection in our setting because in biomedical studies one often collects data on a large number of covariates while a relatively small set of them are believed to be important. Efforts have been made for studies of variable selection in parametric and semiparametric models to cross-sectional data. See Fan and Lv (2010), Fan and Li (2006) and the references therein for a comprehensive survey on the development of variable selection. Recently, variable selection to the longitudinal data framework has been paid great attention (Cantoni, Flemming and Ronchetti (2005); Wang and Qu (2009); Wang et al. (2012); Xue et al. (2010)). Most work in the literature, however, focused on the case where the number of covariates is fixed.

The rest of the article is organized as follows. In Section 2, we introduce the marginal GAPLM model for clustered data analysis, and propose the polynomial spline estimators via a quasi-likelihood approach for the parametric and nonparametric components. We present the asymptotic results of the proposed procedure in Section 3. We study variable selection procedure and develop the associated asymptotic properties in Section 4. Simulation studies and an empirical example are presented in Section 5. Section 6 summarizes the paper and discusses some related issues. Proofs are given in Section 7.

2. Estimation Procedures

For the jth observation of the ith cluster, we observe response variable Y_{ij} and a $(q+p)$-dimensional vector of covariates $(\mathbf{W}_{ij}^{\mathrm{T}}, \mathbf{X}_{ij}^{\mathrm{T}})^{\mathrm{T}}$, where $\mathbf{W}_{ij}$ and $\mathbf{X}_{ij}$ are the nonparametric and parametric components of the marginal GAPLM, respectively, $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$. We assume q, the dimension of $\mathbf{W}_{ij} = (W_{ij1}, \ldots, W_{ijq})^{\mathrm{T}}$, is fixed while p can diverge with the sample size n. Observations from different clusters are independent, but those from the same cluster are correlated. Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^{\mathrm{T}}$ denote the vector of responses for the ith cluster, let $\mathbf{W}_i = (\mathbf{W}_{i1}, \ldots, \mathbf{W}_{im_i})^{\mathrm{T}}$ and $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{im_i})^{\mathrm{T}}$ be the associated $m_i \times q$ and $m_i \times p$ matrices of covariates, respectively.

The GEE approach specifies the first two marginal moments: $\mathrm{E}(Y_{ij}|\mathbf{W}_{ij}, \mathbf{X}_{ij}) =$

$\mu(\theta_{ij})$ and $\text{Var}(Y_{ij}|\mathbf{W}_{ij}, \mathbf{X}_{ij}) = \sigma^2(\theta_{ij}) = \dot{\mu}(\theta_{ij})$, with $\theta_{ij} = \sum_{l=1}^{q} \alpha_l(W_{ijl}) + \mathbf{X}_{ij}^{\text{T}}\boldsymbol{\beta}$. A dispersion parameter can be added in the marginal variance function if overdispersion or underdispersion is suspected to be present. For the marginal logistic regression $\mu(\theta) = e^{\theta}/(1+e^{\theta})$, and for the marginal Poisson regression $\mu(\theta) = e^{\theta}$. The true unknown parameters in the marginal regression model are denoted by $\boldsymbol{\alpha}_0 = \{\alpha_{01}(\cdot), \ldots, \alpha_{0q}(\cdot)\}^{\text{T}}$ and $\boldsymbol{\beta}_0$. In practice, the primary interest is often in the parametric component $\boldsymbol{\beta}_0$. The true value of θ_{ij} is denoted by θ_{0ij}. For simplicity we assume $m_i \equiv m < \infty$.

Without loss of generality, we take the distribution of W_{ijl} to be supported on $[0,1]$. To make the model identifiable, we assume $\int \alpha_l(t)dt = 0, 1 \le l \le q$. We approximate the nonparametric component using B-splines. Let $\tau_0 = 0 < \tau_1 < \cdots < \tau_{K'} < 1 = \tau_{K'+1}$ be a partition of $[0,1]$ into subintervals $[\tau_k, \tau_{k+1}), k = 0, \ldots, K'$ with K' internal knots. A polynomial spline of order J is a function whose restriction to each subinterval is a polynomial of degree $J-1$ and globally $J-2$ times continuously differentiable on $[0,1]$. The collection of splines with a fixed sequence of knots has a B-spline basis $\{B_1(t), \ldots, B_{\tilde{K}}(t)\}$ with $\tilde{K} = K' + J$. Because of the centering constraint $\int \alpha_l = 0, l = 1, \ldots, q$, we focus on the subspace of spline functions $S^0 := \{\phi : \phi = \sum_{k=1}^{\tilde{K}} a_k B_k(t), \int \phi(t) = 0\}$ with normalized basis $\{\sqrt{K}\{B_k(t) - \int B_k(t)dt\}, k = 1, \ldots, K = \tilde{K} - 1\}$. With an abuse of notation, the basis is still denoted by $B_k(t), k = 1, \ldots, K$. This subspace is of dimension $K = \tilde{K} - 1$ due to the zero-integral constraint. Using spline expansions, we can approximate the nonparametric component by $\alpha_l(t) \approx \sum_k a_{lk} B_k(t)$. In our numerical studies, the knots are placed at quantiles of the observed covariate values.

Let $\mathfrak{H}_d$ denote the collection of all functions on the support $[0,1]$ whose uth order derivative satisfies the Hölder condition of order r with $d \equiv u + r$: for each $h \in \mathfrak{H}_d$, there exists a positive constant M_0 such that $|h^{(u)}(s) - h^{(u)}(t)| \le M_0|s-t|^r, \forall\, s, t \in [0,1]$. For B-spline functions (De Boor (2001)), we can find $\mathbf{a}_{0l} = (a_{0l1}, \ldots, a_{0lK})^{\text{T}}$ such that $\| \sum_k a_{0lk} B_k - \alpha_l \|_\infty = O(K^{-d})$ if $\alpha_l \in \mathfrak{H}_d$, where $||\cdot||_\infty$ denotes the supremum norm. Let $\mathbf{a}_0 = (\mathbf{a}_{01}^{\text{T}}, \ldots, \mathbf{a}_{0q}^{\text{T}})^{\text{T}}$, $\mathbf{b}_0 = (\mathbf{a}_0^{\text{T}}, \boldsymbol{\beta}_0^{\text{T}})^{\text{T}}$. Let $\mu_{ij}(\mathbf{b}) = \mu_{ij}(\mathbf{a}, \boldsymbol{\beta}) = \mu(\mathbf{a}^{\text{T}}\mathbf{B}_{ij} + \mathbf{X}_{ij}^{\text{T}}\boldsymbol{\beta})$ with $\mathbf{B}_{ij} = \{B_1(W_{ij1}), \ldots, B_K(W_{ij1}), \ldots, B_K(W_{ijq})\}^{\text{T}} \in R^{qK}$, and similarly we define $\sigma_{ij}^2(\mathbf{b})$. Let $\boldsymbol{\mu}_i(\mathbf{b}) = \{\mu_{i1}(\mathbf{b}), \ldots, \mu_{im}(\mathbf{b})\}^{\text{T}}$, $\boldsymbol{\mu}_{0i} = \{\mu(\theta_{0i1}), \ldots, \mu(\theta_{0im})\}^{\text{T}}$, $\mathbf{A}_i(\mathbf{b}) = \text{diag}\{\sigma_{i1}^2(\mathbf{b}), \ldots, \sigma_{im}^2(\mathbf{b})\}$, and $\mathbf{A}_{0i} = \text{diag}\{\sigma^2(\theta_{0i1}), \ldots, \sigma^2(\theta_{0im})\}$. For a vector $\mathbf{a}$, $\|\mathbf{a}\|$ denotes its l_2 (Euclidean) norm; and for a matrix $\mathbf{A}$, $\|\mathbf{A}\|$ denotes its

Frobenius norm.

Let $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{im})^{\intercal}_{m\times p}$ and $\mathbf{U}_i = \{(\mathbf{B}_{i1}, \ldots, \mathbf{B}_{im})^{\intercal}, \mathbf{X}_i\}_{m\times(qK+p)}$, and let $\mathbf{U}_{ij}$ be the j-th row of $\mathbf{U}_i$. The *GEE estimator* $\widehat{\mathbf{b}}$ is the solution of

$$\mathbf{S}(\mathbf{b}) = \sum_{i=1}^{n} \mathbf{U}_i^{\intercal}\mathbf{A}_i^{1/2}(\mathbf{b})\widehat{\mathbf{R}}^{-1}\mathbf{A}_i^{-1/2}(\mathbf{b})\{\mathbf{Y}_i - \boldsymbol{\mu}_i(\mathbf{b})\} = 0, \tag{2.3}$$

where $\widehat{\mathbf{R}}$ is an estimate of the working correlation matrix. Some commonly used working correlation structures include independence, AR-1, equally correlated (also called compound symmetry), and unstructured correlation, among others.

The theory established in this paper does not require that $\widehat{\mathbf{R}}$ consistently estimates the true common correlation matrix $\mathbf{R}_0$, although the deviation from $\mathbf{R}_0$ may affect the efficiency of the GEE estimator. We introduce a residual-based estimator $\widehat{\mathbf{R}}$ for the unstructured working correlation matrix when an initial estimator $\widetilde{\mathbf{b}}$ is available. A simple way to obtain the initial estimator is to solve the GEE under the working independence assumption,

$$\widetilde{\mathbf{S}}(\mathbf{b}) = \sum_{i=1}^{n} \mathbf{U}_i^{\intercal}\{\mathbf{Y}_i - \boldsymbol{\mu}_i(\mathbf{b})\} = 0. \tag{2.4}$$

Then we use a moment estimator to estimate the unstructured correlation matrix:

$$\widehat{\mathbf{R}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{A}_i^{-1/2}(\widetilde{\mathbf{b}})\{\mathbf{Y}_i - \boldsymbol{\mu}_i(\widetilde{\mathbf{b}})\}\{\mathbf{Y}_i - \boldsymbol{\mu}_i(\widetilde{\mathbf{b}})\}^{\intercal}\mathbf{A}_i^{-1/2}(\widetilde{\mathbf{b}}). \tag{2.5}$$

3. Main Theoretical Results

In this section, we investigate the asymptotic properties of the estimator in (2.3) when the dimension of $\mathbf{X}_{ij}$ increases with the number of clusters n at an appropriate rate. To facilitate the presentation, we fix some regularity conditions. Let $r_n = \sqrt{(K+p)/n + K^{-2d}}$.

(A1) $\sup_{i,j} ||\mathbf{X}_{ij}|| = O(\sqrt{p})$.

(A2) $\alpha_l \in \mathfrak{H}_d$ for some $d > 1/2$, $l = 1, \ldots, q$.

(A3) There exist a finite constant $M_1 > 0$ such that $E(||\mathbf{Y} - \boldsymbol{\mu}_0||^{2+\delta}) \leq M_1$ for some $\delta > 0$.

(A4) There exist positive constants c_1 and c_2 such that

$$c_1 \leq \lambda_{\min}\left(n^{-1}\sum_{i=1}^{n}\mathbf{U}_i^{\intercal}\mathbf{U}_i\right) \leq \lambda_{\max}\left(n^{-1}\sum_{i=1}^{n}\mathbf{U}_i^{\intercal}\mathbf{U}_i\right) \leq c_2,$$

where $\lambda_{\min}$ ($\lambda_{\max}$) denotes the smallest (largest) eigenvalue of a matrix.

(A5) Let $\mathbf{C} = \{\mathbf{b} : ||\mathbf{b} - \mathbf{b}_0|| \leq \Delta r_n\}$, where Δ is a finite positive constant. Then $\dot{\mu}(\mathbf{U}_{ij}^{\intercal}\mathbf{b})$, $1 \leq i \leq n, 1 \leq j \leq m$, is uniformly bounded away from 0 and ∞ on $\mathbf{C}$; $\ddot{\mu}(\mathbf{U}_{ij}^{\intercal}\mathbf{b})$ and $\mu^{(3)}(\mathbf{U}_{ij}^{\intercal}\mathbf{b})$, $1 \leq i \leq n, 1 \leq j \leq m$, are uniformly bounded by a finite positive constant M_2 on $\mathbf{C}$.

Remark 3.1. Condition (A1) is a common assumption in the literature on M-estimation with diverging dimension. For a B-spline basis (De Boor (2001)), $\|\mathbf{B}_{ij}\| = O(\sqrt{K})$ and thus (A1) implies that $\sup_{i,j}\|\mathbf{U}_{ij}\| = O(\sqrt{K+p})$. When each cluster has only one observation, (A4) is commonly imposed on semiparametric regression for independent data. Condition (A4) implies the eigenvalues of $\sum_i \mathbf{X}_i^{\intercal}\mathbf{X}_i/n$ are bounded away from zero and infinity. Conditions (A3) and (A5) were also assumed in Wang (2011).

The following proposition demonstrates the consistency of the initial estimator defined in (2.4) and the estimated working correlation matrix defined in (2.5).

Proposition 3.1. *Under (A1)–(A5) and $p/n \to 0$, $K\log K/n \to 0, K \to \infty$,*

(i) the GEE in (2.4) has a root $\widetilde{\mathbf{b}}$ satisfying $\|\widetilde{\mathbf{b}} - \mathbf{b}_0\| = O_p(r_n)$;

(ii) the estimated correlation matrix satisfies $\|\widehat{\mathbf{R}} - \mathbf{R}_0\| = O_p(r_n)$ and $\|\widehat{\mathbf{R}}^{-1} - \mathbf{R}_0^{-1}\| = O_p(r_n)$.

Based on this proposition, it is natural to make the following assumption.

(A6) The common true correlation matrix $\mathbf{R}_0$ has eigenvalues bounded away from zero and $+\infty$. The estimated working correlation matrix $\widehat{\mathbf{R}}$ satisfies $||\widehat{\mathbf{R}}^{-1} - \overline{\mathbf{R}}^{-1}|| = O_p(r_n)$, where $\overline{\mathbf{R}}$ is a constant positive-definite matrix with eigenvalues bounded away from zero and $+\infty$. It is worth noting that we do not require $\overline{\mathbf{R}}$ to be the true correlation matrix $\mathbf{R}_0$.

Proposition 3.1 guarantees that (A6) is satisfied when a nonparametric moment estimator is used for the working correlation matrix, with $\overline{\mathbf{R}} = \mathbf{R}_0$. However, we allow $\overline{\mathbf{R}} \neq \mathbf{R}_0$ as what would happen when a misspecified parametric model is used to estimate the correlation matrix.

Take

$$\overline{\mathbf{S}}(\mathbf{b}) = \sum_{i=1}^{n} \mathbf{U}_i^{\intercal} \mathbf{A}_i^{1/2}(\mathbf{b}) \overline{\mathbf{R}}^{-1} \mathbf{A}_i^{-1/2}(\mathbf{b}) \{\mathbf{Y}_i - \boldsymbol{\mu}_i(\mathbf{b})\}$$

as an approximation of $\mathbf{S}(\mathbf{b})$ with estimated correlation matrix replaced by its asymptotic limit. Let

$$\overline{\mathbf{M}}(\mathbf{b}) = \sum_{i=1}^{n} \mathbf{U}_i^{\intercal} \mathbf{A}_i^{1/2}(\mathbf{b}) \overline{\mathbf{R}}^{-1} \mathbf{R}_0 \overline{\mathbf{R}}^{-1} \mathbf{A}_i^{1/2}(\mathbf{b}) \mathbf{U}_i$$

be the covariance matrix of $\overline{\mathbf{S}}(\mathbf{b})$. Furthermore, let $\mathbf{D}(\mathbf{b}) = -\frac{\partial}{\partial \mathbf{b}^{\intercal}} \mathbf{S}(\mathbf{b})$ and $\overline{\mathbf{D}}(\mathbf{b}) = -\frac{\partial}{\partial \mathbf{b}^{\intercal}} \overline{\mathbf{S}}(\mathbf{b})$.

As in Wang (2011), $\overline{\mathbf{D}}(\mathbf{b})$ can be decomposed as

$$\overline{\mathbf{D}}(\mathbf{b}) = \overline{\mathbf{H}}(\mathbf{b}) + \overline{\mathbf{E}}(\mathbf{b}) + \overline{\mathbf{G}}(\mathbf{b}), \tag{3.4}$$

where the first term $\overline{\mathbf{H}}(\mathbf{b})$ will be shown to dominate the other two in our analysis, with

$$\begin{aligned}
\overline{\mathbf{H}}(\mathbf{b}) &= \sum_{i=1}^{n} \mathbf{U}_i^{\intercal} \mathbf{A}_i^{1/2}(\mathbf{b}) \overline{\mathbf{R}}^{-1} \mathbf{A}_i^{1/2}(\mathbf{b}) \mathbf{U}_i, \\
\overline{\mathbf{E}}(\mathbf{b}) &= \frac{1}{2} \sum_{i=1}^{n} \mathbf{U}_i^{\intercal} \mathbf{A}_i^{1/2}(\mathbf{b}) \overline{\mathbf{R}}^{-1} \mathbf{A}_i^{-3/2}(\mathbf{b}) \mathbf{C}_i(\mathbf{b}) \mathbf{F}_i(\mathbf{b}) \mathbf{U}_i, \\
\overline{\mathbf{G}}(\mathbf{b}) &= -\frac{1}{2} \sum_{i=1}^{n} \mathbf{U}_i^{\intercal} \mathbf{A}_i^{1/2}(\mathbf{b}) \mathbf{F}_i(\mathbf{b}) \mathbf{J}_i(\mathbf{b}) \mathbf{U}_i,
\end{aligned}$$

with $\mathbf{C}_i(\mathbf{b}) = \text{diag}\{Y_{i1} - \mu_{i1}(\mathbf{b}), \ldots, Y_{im} - \mu_{im}(\mathbf{b})\}$, $\mathbf{F}_i(\mathbf{b}) = \text{diag}\{\ddot{\mu}(\mathbf{U}_{i1}^{\intercal}\mathbf{b}), \ldots, \ddot{\mu}(\mathbf{U}_{im}^{\intercal}\mathbf{b})\}$, $\mathbf{J}_i(\mathbf{b}) = \text{diag}[\overline{\mathbf{R}}^{-1} \mathbf{A}_i^{-1/2}(\mathbf{b}) \{\mathbf{Y}_i - \boldsymbol{\mu}_i(\mathbf{b})\}]$.

The following theorem provides the existence and the convergence rates of the semiparametric GEE estimator.

Theorem 3.1. *(*Existence and consistency*). Suppose (A1)–(A6) hold and that* $p^2/n \to 0$*,* $K^2/n \to 0, p/K^{2d} \to 0$*. Then* $\mathbf{S}(\mathbf{b}) = 0$ *has a root* $\widehat{\mathbf{b}}$ *satisfying*

$$||\widehat{\mathbf{b}} - \mathbf{b}_0|| = O_p(r_n).$$

As an immediate implication, with $\widehat{\alpha}_l = \sum_{k=1}^{K} \hat{a}_{lk} B_k$,

$$\sum_{l=1}^{q} \|\widehat{\alpha}_l - \alpha_{0l}\|^2 + \sum_{j=1}^{p} |\hat{\beta}_j - \beta_{0j}|^2 = O_p(r_n^2).$$

The parametric part can be shown to be asymptotically normal under slightly stronger conditions. Let

$$\mathcal{F}_j := \Bigg\{ g : g(\mathbf{W}_{1j}) = \sum_{l=1}^{q} h_l(W_{1jl}) \text{ for some functions } h_j \text{ with } \int h_l = 0 \text{ and } E\sum_{l=1}^{q} h_l^2(W_{1jl}) < \infty \Bigg\},$$

and let $\mathcal{F}^m = \{(g_1, \ldots, g_m) : g_j \in \mathcal{F}_j\}$ be the Cartesian product of $\mathcal{F}_j$. For any random vector $\Omega \in R^m$ with $E(\|\Omega\|^2) < \infty$, let $E_{\mathcal{F}^m}(\Omega)$ denote the projection of Ω onto $\mathcal{F}^m$ in the sense that

$$\begin{aligned} &E[\{\Omega - E_{\mathcal{F}^m}(\Omega)\}^{\intercal} \mathbf{A}_{01}^{1/2} \overline{\mathbf{R}}^{-1} \mathbf{A}_{01}^{1/2} \{\Omega - E_{\mathcal{F}^m}(\Omega)\}] \\ &\quad = \inf_{(g_1,\ldots,g_m)^{\intercal} \in \mathcal{F}^m} E\{(\Omega - \mathbf{g})^{\intercal} \mathbf{A}_{01}^{1/2} \overline{\mathbf{R}}^{-1} \mathbf{A}_{01}^{1/2} (\Omega - \mathbf{g})\}. \end{aligned}$$

where $\mathbf{g} = \{g_1(\mathbf{W}_{11}), \ldots, g_m(\mathbf{W}_{1m})\}^{\intercal}$. When Ω is an $m \times p$ matrix, we take $E_{\mathcal{F}^m}(\Omega)$ to be the $m \times p$ matrix whose columns are the projections of the columns of Ω. This extends the setup in Li (2000) to the longitudinal data framework.

Let Γ be the projection of $\mathbf{X}_1$ onto $\mathcal{F}^m$, so $\Gamma(\mathbf{W}_i)$ is a $m \times p$ matrix whose (j, s)-entry can be written as $\sum_{l=1}^{q} h_{jl}^{(s)}(W_{ijl})$.

(A7) $h_{jl}^{(s)} \in \mathfrak{H}_d, 1 \le j \le m, 1 \le l \le q, 1 \le s \le p$.

Theorem 3.2. (Asymptotic normality) *Suppose Conditions (A1)–(A7) hold. If* $p^3/n \to 0, p/K^{2d-1} \to 0, p^2/K^{2d} \to 0, np/K^{4d} \to 0$ *and* $pK^2/n \to 0$*, then for any unit vector* $\boldsymbol{\alpha} \in R^p$*,*

$$\boldsymbol{\alpha}^{\intercal} \overline{\mathcal{M}}_0^{-1/2} \overline{\mathcal{H}}_0 (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to N(0, 1),$$

in distribution, where $\overline{\mathcal{M}}_0 = \sum_i \{\mathbf{X}_i - \Gamma(\mathbf{W}_i)\}^{\intercal} \mathbf{A}_{0i}^{1/2} \overline{\mathbf{R}}^{-1} \mathbf{R}_0 \overline{\mathbf{R}}^{-1} \mathbf{A}_{0i}^{1/2} \{\mathbf{X}_i - \Gamma(\mathbf{W}_i)\}$ *and* $\overline{\mathcal{H}}_0 = \sum_i \{\mathbf{X}_i - \Gamma(\mathbf{W}_i)\}^{\intercal} \mathbf{A}_{0i}^{1/2} \overline{\mathbf{R}}^{-1} \mathbf{A}_{0i}^{1/2} \{\mathbf{X}_i - \Gamma(\mathbf{W}_i)\}$.

Remark 3.2. Theorem 3.2 suggests that the covariance matrix of $\widehat{\boldsymbol{\beta}}$ is approximately $\boldsymbol{\Sigma} = \overline{\mathcal{H}}_0^{-1} \overline{\mathcal{M}}_0 \overline{\mathcal{H}}_0^{-1}$. To estimate $\boldsymbol{\Sigma}$, we can use the *sandwich covariance matrix estimator*

$$\widehat{\boldsymbol{\Sigma}} = \mathcal{H}^{-1}(\widehat{\mathbf{b}}) \widehat{\mathcal{M}}(\widehat{\mathbf{b}}) \mathcal{H}^{-1}(\widehat{\mathbf{b}}),$$

where $\widehat{\mathcal{M}}(\mathbf{b}) = \sum_i \{\mathbf{X}_i - \widehat{\Gamma}(\mathbf{W}_i)\}^{\mathrm{T}} \mathbf{A}_i^{1/2}(\mathbf{b}) \widehat{\mathbf{R}}^{-1} \mathbf{R}_0 \widehat{\mathbf{R}}^{-1} \mathbf{A}_i^{1/2}(\mathbf{b}) \{\mathbf{X}_i - \widehat{\Gamma}(\mathbf{W}_i)\}$ and $\mathcal{H}(\mathbf{b}) = \sum_i \{\mathbf{X}_i - \widehat{\Gamma}(\mathbf{W}_i)\}^{\mathrm{T}} \mathbf{A}_i^{1/2}(\mathbf{b}) \widehat{\mathbf{R}}^{-1} \mathbf{A}_i^{1/2}(\mathbf{b}) \{\mathbf{X}_i - \widehat{\Gamma}(\mathbf{W}_i)\}$. The unknown true covariance matrix $\mathbf{R}_0$ can be estimated from an initial estimator as in Proposition 3.1 under the unstructured working correlation. The sandwich estimator can be shown to be consistent by standard arguments.

4. Variable Selection

We consider variable selection for the covariates that appear in the linear part. We assume the the linear part in the true model is sparse in the sense that the majority of the components of $\boldsymbol{\beta}$ are zero. Without loss of generality, we assume that only the first s coefficients of $\boldsymbol{\beta}$ are nonzero.

We consider simultaneous variable selection and estimation based on the penalized GEE:

$$\mathbf{L}(\mathbf{b}) = \mathbf{S}(\mathbf{b}) - n\mathbf{q}_\lambda(|\boldsymbol{\beta}|)\mathrm{sgn}(\boldsymbol{\beta}), \tag{4.5}$$

where $\mathbf{S}(\mathbf{b})$ is the estimating function in (2.3), $\mathbf{q}(|\boldsymbol{\beta}|) = (\mathbf{0}_{qk}^T, q_\lambda(|\beta_1|), \ldots, q_\lambda(|\beta_p|))^T$, $\mathrm{sgn}(\boldsymbol{\beta}) = (\mathbf{0}_{\mathrm{qk}}^{\mathrm{T}}, \mathrm{sgn}(\beta_1), \ldots, \mathrm{sgn}(\beta_{\mathrm{p}}))^{\mathrm{T}}$ with $\mathbf{0}_{qk}$ denoting a qk-dimensional vector of zeros, $\mathrm{sgn}(t) = I(t > 0) - I(t < 0)$ being the sign function, and $\mathbf{q}_\lambda(|\boldsymbol{\beta}|)\mathrm{sgn}(\boldsymbol{\beta})$ is the componentwise product of $\mathbf{q}_\lambda(|\boldsymbol{\beta}|)$ and $\mathrm{sgn}(\boldsymbol{\beta})$. In general the penalty $q_\lambda(|\beta_j|)$ should be close to zero when $|\beta_j|$ is large so that little extra bias is introduced by the penalty term. On the other hand, $q_\lambda(|\beta_j|)$ should be large if $|\beta_j|$ is close to zero to encourage it to be shrunk to zero. Different penalty functions have been proposed. Here we focus on the smoothly clipped absolute deviation (SCAD) penalty of Fan and Li (2001),

$$q_\lambda(x) = \lambda \left\{ I(|x| \le \lambda) + \frac{(c\lambda - |x|)_+}{(c-1)\lambda} I(|x| > \lambda) \right\} \quad \text{for some} \quad c > 2,$$

where the notation $(t)_+$ stands for the positive part of t and λ is a tuning parameter that determines the level of penalty. Fan and Li (2001) suggested using $c = 3.7$ for the SCAD penalty function.

The following additional condition is adopted.

(A8) $\min_{1 \le j \le s} |\beta_{0j}|/\lambda \to \infty$, $s^3/n \to 0$, $s/K^{2d-1} \to 0$, $s^2/K^{2d} \to 0$, $ns/K^{4d} \to 0$, $sK^2/n \to 0$, $\lambda \to 0$, $s^2(\log n)^4 = o(n\lambda^2)$, $\log p = o(n\lambda^2/(\log n)^2)$, $ps^4(\log n)^6 = o(n^2\lambda^2)$ and $ps^3(\log n)^8 = o(n^2\lambda^4)$.

The first part of (A8) indicates that the smallest signal does not converge to zero too fast so that the nonzero coefficients can be distinguished from the zero ones asymptotically. If s is fixed and $\min_{1\le j\le s}|\beta_{0j}|$ is bounded away from zero, $p = n$ is allowed in Condition (A8).

Since $\mathbf{L}$ is not continuous, an exact solution to $\mathbf{L}(\mathbf{b}) = 0$ may not exist. An exact solution is replaced by the zero-crossing condition (Johnson, Lin and Zeng (2008)), which roughly means a small perturbation of any zero component in $\hat{\boldsymbol{\beta}}$ changes the sign of the penalized estimating equations. We still use $\hat{\mathbf{b}}$ to denote the approximate solution here.

Theorem 4.3. *Under (A1)-(A8), there exists $\hat{\mathbf{b}} = (\hat{\mathbf{a}}^{\mathrm{T}}, \hat{\boldsymbol{\beta}}^{\mathrm{T}})^{\mathrm{T}}$ that satisfies the following*

(i) $P(\hat{\boldsymbol{\beta}}^{(2)} = 0) \to 1$, *where* $\hat{\boldsymbol{\beta}}^{(2)} = (\hat{\beta}_{s+1}, \ldots, \hat{\beta}_p)^T$.

(ii) $\mathbf{L}_j(\hat{\mathbf{b}}) = 0$ *for* $1 \le j \le qK + s$, *where* $\mathbf{L}_j$ *is the* j*th component of the* $(qK + p)$*-dimensional* $\mathbf{L}$.

(iii) $\overline{\lim}_{n\to\infty}\overline{\lim}_{\epsilon\to 0+} n^{-1}\mathbf{L}_j(\hat{\mathbf{b}}+\epsilon\mathbf{e}_j)\mathbf{L}_j(\hat{\mathbf{b}}-\epsilon\mathbf{e}_j) \le 0$, *for* $qK+s+1 \le j \le qK+p$, *where* $\mathbf{e}_j$ *is the* $(qK + p)$*-dimensional vector with all components zero except for a one at position* j.

(iv) The convergence rate of the penalized estimator is $O_p(r_n)$, *as in Theorem 3.1.*

(v) With $\hat{\boldsymbol{\beta}}^{(1)} = (\hat{\beta}_1, \ldots, \hat{\beta}_s)^T$,

$$\boldsymbol{\alpha}^{\mathrm{T}}\overline{\mathcal{M}}_{01}\overline{\mathcal{H}}_{01}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to N(0, 1)$$

in distribution, where $\overline{\mathcal{M}}_{01}$ *and* $\overline{\mathcal{H}}_{01}$ *are the principal submatrices of* $\overline{\mathcal{M}}_0^{-1/2}$ *and* $\overline{\mathcal{H}}_0$ *respectively, by removing the last* $p - s$ *columns and rows.*

Property (i) implies model selection consistency. Property (ii) shows that $\hat{\mathbf{b}}$ is an exact solution for the first $qK + s$ equations in $\mathbf{L}(\mathbf{b}) = 0$, and that in particular the penalty has no effect for these equations. Property (iii) is the approximate zero-crossing property.

Remark 4.3. The penalized GEE in (4.5) can be efficiently solved by combining the MM algorithm (Hunter and Li (2005)) with the Newton-Raphson algorithm, as in Wang et al. (2012). Motivated by the recent work of Chen and Chen (2008), we proposed a

high-dimensional BIC (HBIC) criterion under the working independence assumption to compare the estimators from the solution path:

$$\text{HBIC}(\lambda) = \log(\widehat{\sigma}^2_\lambda) + |M_\lambda| \frac{C_n \log(p)}{n},$$

where $|M_\lambda|$ denotes the cardinality of the model selected when the tuning parameter λ is used, and $\widehat{\sigma}^2_\lambda = n^{-1}\text{SSE}_\lambda$ with $\text{SSE}_\lambda = ||\mathbf{Y} - \boldsymbol{\mu}_i(\hat{\mathbf{b}})||^2$ with $\hat{\mathbf{b}}$ the penalized estimator corresponding to the tuning parameter λ. As we are interested in the case where p grows with n, the penalty term also depends on p, and C_n is a sequence of numbers that diverges to ∞, which we take to be $\log(n)$ in our numerical studies in Section 5. We choose the value of λ that minimizes $\text{HBIC}(\lambda)$. Although HBIC performs satisfactorily in our numerical studies, it is a challenge to establish the relevant theory in the GEE setting for models with partially linear structure.

5. Numerical Studies

In this section, we carry out Monte-Carlo studies to assess the numerical performance of estimation in the semiparametric marginal regression model and the proposed variable selection procedure. We then apply the proposed methods to analyze a data set from the Wisconsin Epidemiological Study of Diabetic Retinopathy.

Example 1 (continuous response). The correlated normal responses were generated from the model

$$Y_{ij} = \alpha_1(W_{ij1}) + \alpha_2(W_{ij2}) + \mathbf{X}_{ij}^{\text{T}}\boldsymbol{\beta} + \epsilon_{ij},$$

where $i = 1, ..., 200$, $j = 1, ..., 4$, $\boldsymbol{\beta}$ is a p_n-dimensional vector of parameters with $p_n = \lfloor 2.5n^{1/3} \rfloor$, with $\lfloor q \rfloor$ denoting the largest integer not greater than q. In this example, $\boldsymbol{\beta}^{\text{T}} = (1 \cdot \mathbf{1}_k^{\text{T}}, -1.5 \cdot \mathbf{1}_k^{\text{T}}, 1.8 \cdot \mathbf{1}_{p_n-2k}^{\text{T}})$, where $\mathbf{1}_k$ denotes a k-dimensional vector of ones and $k = \lfloor p_n/3 \rfloor$. The nonparametric components were $\alpha_1(t) = \exp(-t) - \exp(-0.5)$ and $\alpha_2(t) = \sin\{4 * (t - 0.5)\}$. In addition, $\mathbf{X}_{ij} = (x_{ij1}, \ldots, x_{ijp_n})^{\text{T}}$ had a multivariate normal distribution with mean zero, marginal variance 0.2 and an AR-1 correlation matrix with autocorrelation coefficient 0.5. The covariates W_{ij1} and W_{ij2} were independent uniform distributed on [0,1] and independent of $\mathbf{X}_{ij}$ and ϵ_{ij}. The random error $(\epsilon_{i1}, ..., \epsilon_{i4})^{\text{T}}$ was generated from the multivariate normal distribution with marginal mean 0, marginal variance 1 and an AR-1 correlation matrix with autocorrelation coefficient 0.5.

For each setup, we generated 400 data sets. We evaluated the accuracy of the estimators for the regression parameter $\boldsymbol{\beta}$ by MSE $= 400^{-1}\sum_{k=1}^{400}||\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}||^2$, where $\widehat{\boldsymbol{\beta}}^{(k)}$ denotes the estimated value of $\boldsymbol{\beta}$ in the kth simulation run. To evaluate the estimation for $\alpha_l(\cdot)$, we used

$$\text{ADE}_l = \frac{1}{400mn}\sum_{k=1}^{400}\sum_{i=1}^{n}\sum_{j=1}^{m}|\widehat{\alpha}_l^{(k)}(W_{ijl})-\alpha_l(W_{ijl})|,$$

where $\widehat{\alpha}_l^{(k)}(W_{ijl})$ is the spline approximation to $\alpha_l(W_{ijl})$ in the kth simulation run. In the simulations, we used cubic B-splines with six degrees of freedom and took three quartiles of the W_{ijk}, $k=1,2$, as internal knots.

In Table 5.1, we summarize the MSE for estimating $\boldsymbol{\beta}$ and the ADE for estimating $\alpha_1(\cdot)$ and $\alpha_2(\cdot)$, respectively. We consider four working correlation structures: independence(INDE), exchangeable (EXCH), AR(1) (AR1), and unstructured working correlation (UNST). We observe that even when the covariate dimension grows at an appropriate rate with the sample size, the accuracy of the proposed GEE estimator for both the regression parameters and the nonparametric functions is satisfactory. In the simulations, we were not aiming to produce optimal estimators for the nonparametric components. Instead, we focused on estimating the parametric components, which only requires consistent estimators for the nonparametric components. We note that the independent working correlation structures shows worse performance in estimating the parameters in the linear part.

Figure 5.1 depicts the nonparametric components $\alpha_1(t)$ and $\alpha_2(t)$. Also imposed on the graph are the estimators using each of the working correlation structures based on one randomly simulated data set.

Example 2 (binary response). We considered the following model for the marginal expectation of Y_{ij},

$$\text{logit}(E(Y_{ij}|\mathbf{X}_{ij},W_{ij1},W_{ij2})) = \alpha_1(W_{ij1})+\alpha_2(W_{ij2})+\mathbf{X}_{ij}^{\text{T}}\boldsymbol{\beta}, \tag{5.6}$$

where $i=1,...,400$, $j=1,...,3$, $\boldsymbol{\beta}^{\text{T}}=(0.4\cdot\mathbf{1}_k^{\text{T}},-0.3\cdot\mathbf{1}_k^{\text{T}},0.2\cdot\mathbf{1}_k^{\text{T}},-0.1\cdot\mathbf{1}_{p_n-3k}^{\text{T}})$, where $p_n=\lfloor 2.5n^{1/3}\rfloor$ and $k=\lfloor p_n/4\rfloor$. The nonparametric components were $\alpha_1(t)=\exp(-t)-\exp(-0.5)$ and $\alpha_2(t)=(t-0.5)^3$. The distributions of $\mathbf{X}_{ij}$, W_{ij1} and W_{ij2} were the same as those in Example 1. The binary response vector for each cluster had

Table 5.1: Example 1: Longitudinal data with continuous responses. ADE: absolute deviation error; INDE: independence; EXCH: exchangeable; AR1: AR(1); and UNST: unstructured.

n	p_n		MSE	ADE_1	ADE_2
200	14	INDE	0.1552	0.0808	0.1264
		EXCH	0.1216	0.0741	0.1287
		AR1	0.1030	0.0693	0.1297
		UNST	0.1053	0.0703	0.1304
1000	24	INDE	0.0517	0.0493	0.0989
		EXCH	0.0405	0.0473	0.1059
		AR1	0.0350	0.0448	0.1092
		UNST	0.0352	0.0449	0.1091
2000	31	INDE	0.0327	0.0437	0.0947
		EXCH	0.0257	0.0424	0.1017
		AR1	0.0222	0.0402	0.1052
		UNST	0.0222	0.0402	0.1051

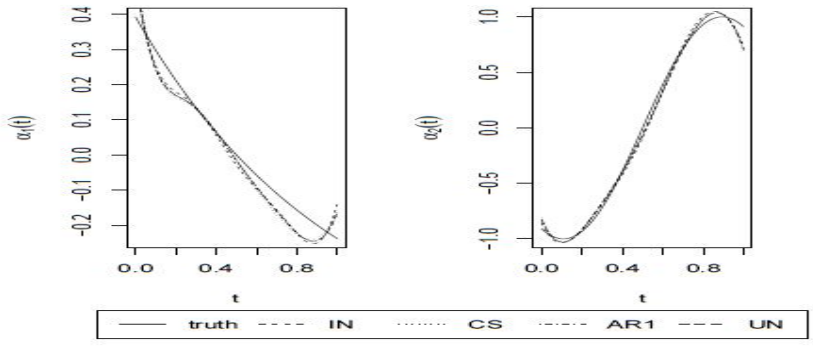

Figure 5.1: Plot of the nonparametric components and their estimates using each of the working correlation structures based on one random sample.

Table 5.2: Example 2: Longitudinal data with binary responses.

n	p_n		MSE	ADE_1	ADE_2
400	18	INDE	0.5870	0.1347	0.1382
		EXCH	0.3512	0.1054	0.1140
		AR1	0.4164	0.1142	0.1234
		UNST	0.3563	0.1046	0.1124
1000	24	INDE	0.3587	0.0907	0.0899
		EXCH	0.2531	0.0779	0.0779
		AR1	0.2780	0.0823	0.0838
		UNST	0.2607	0.0782	0.0777
2000	34	INDE	0.2637	0.0730	0.0721
		EXCH	0.2138	0.0657	0.0665
		AR1	0.2255	0.0689	0.0700
		UNST	0.2224	0.0659	0.0663

the above marginal mean and an exchangeable correlation structure with correlation coefficient 0.5. Such correlated binary data were generated using Bahadur's representation (see, for example, Fitzmaurice (1995)).

Table 2 summarizes the results for this example based on 400 simulation runs. As with Example 1, Example 2 also verifies that for binary responses the GEE estimator is satisfactory when the covariate dimension grows, and that more efficient estimation is achieved when the true correlation matrix is used.

Example 3 (variable selection with binary response). We generated correlated binary data from model (5.6) with $n = 400$, $m = 3$, $\alpha_1(t) = \exp(-t) - \exp(-0.5)$ and $\alpha_2(t) = \sin\{4 * (t - 0.5)\}$. We considered $p_n = 50$ and 100. The true value for $\boldsymbol{\beta}^{\text{T}}$ was $(0.7, -0.7, -0.7, 0.7, -0.7, 0.7, -0.7, 0.7, -0.7, 0.7, \mathbf{0}_{p_n-10})$. Thus the number of nonzero coefficients was $s = 10$. The true correlation structure was exchangeable with correlation coefficient 0.5.

We report the results based on 400 simulation runs in Table 3 for four scenarios with working correlation structures INDE, EXCH, AR1, and UNST. We use the estimator from the unpenalized GEE as the initial value. The algorithm stops if $||\widehat{\boldsymbol{\beta}}^{(k+1)} - \widehat{\boldsymbol{\beta}}^{(k)}|| \leq 10^{-3}$. The results in Table 3 include (i) the average true positives (TP); that is, the average number of selected covariates that correspond to the nonzero coefficients in the

Table 5.3: Example 3: Variable selection for longitudinal data with binary responses.

p_n		TP	FP	EXACT	L_2 error
50	INDE	8.8625	0.5725	0.3325	0.8126
	EXCH	8.9950	0.1825	0.5025	0.7162
	AR1	8.9575	0.2400	0.4525	0.7537
	UNST	9.0175	0.1975	0.5075	0.7537
100	INDE	8.8175	1.3600	0.2025	0.8391
	EXCH	8.9400	0.3750	0.3975	0.7412
	AR1	8.8425	0.6225	0.3200	0.8089
	UNST	8.9325	0.3725	0.4025	0.8089

underlying model; (ii) the average false positives (FP), the average number of selected covariates that correspond to the zero coefficients in the underlying model; (iii) the percentage of times the true model is exactly selected (EXACT); and (iv) the squared L_2 estimation error for estimating the nonzero coefficients. We observe that in all scenarios, the proposed variable selection procedure is able to pick a model which has true positives close to $s = 10$ and rather few false negatives. The L_2 estimation error is also very small. The best performance is obtained when the exchangeable working correlation structure or the unstructured working correlation is used.

Example 4 (analysis of diabetic retinopathy data). The data set is from the Wisconsin Epidemiological Study of Diabetic Retinopathy (Klein et al. (1984)). The binary response variable indicates the presence or absence of diabetic retinopathy in each of the two eyes from each of the 720 individuals in the study. There are 13 potential risk factors: X_1 (eye refractive error), X_2 (eye intraocular pressure), X_3 (age at diagnosis of diabetes), X_4 (duration of diabetes), X_5 (glycosylated hemoglobin level), X_6 (systolic blood pressure), X_7 (diastolic blood pressure), X_8 (body mass index), X_9 (pulse rate), X_{10} (gender, male=1, female=2), X_{11} (proteinuria, absent=0, present=1), X_{12} (doses of insulin per day), and X_{13} (residence, urban=0, rural=1). Barnhart and Williamson (1998) analyzed this data set using GEE with quadratic effects in X_4 and X_8. Based on preliminary data exploration, we applied GEE with marginal GAPLM and the logit link function. More specifically, we modeled the effects of X_4 and X_8 using cubic splines and included the other covariates (and the 21 first-order interactions among the contin-

Table 5.4: Analysis of diabetic retinopathy data: results for the parametric part using two working correlation structures (INDE and UNST). The numbers reported are the GEE estimators (the numbers in the parenthesis are the associated standard errors).

	INDE	UNST
X_1	-0.6177(0.2573)	-0.6302(0.2428)
X_2	-0.0476(0.1881)	-0.0281(0.1720)
X_5	-0.8010(0.3417)	-0.8094(0.3415)
X_7	-0.0552(0.0721)	-0.0567(0.0715)
X_9	0.2169(0.0918)	0.2077(0.0910)
X_1*X_2	0.0212(0.0106)	0.0209(0.0094)
X_1*X_5	0.0224(0.0162)	0.0237(0.0158)
X_2*X_7	0.0015(0.0024)	0.0011(0.0022)
X_5*X_7	0.0125(0.0046)	0.0126(0.0046)
X_7*X_9	-0.0025(0.0012)	-0.0024(0.0011)

uous covariates) in the linear part. This resulted in a model with 33 terms in the linear part.

We first applied our variable selection procedure to select the variables in the linear part. The following covariates were selected: X_1, X_2, X_5, X_1*X_2, X_1*X_5, X_2*X_7, X_5*X_7, X_7*X_9. We refit the model with the selected variables in the linear part (X_7 and X_9 were also included) and X_4, X_8 in the nonlinear part. Table 5.4 summarizes the estimates and the standard errors (computed from the sandwich covariance formula) for the parametric part of the marginal GAPLM using the independence and unstructured correlation structures. Here, there are just two observations within each cluster. Thus the unstructured correlation is equivalent to the exchangeable or AR(1) structure. In Figure 5.2, we depict the estimated nonparametric functions for X_4 and X_8 under the unstructured working correlation. Some risk factors selected by our approach, such as glycosylated hemoglobin level, diastolic blood pressure, were also identified as important by Barnhart and Williamson (1998). However, they did not consider interaction effects and they modeled both the effect of the duration of diabetes and that of body mass index as quadratic. Our estimation of the nonparametric components suggests that the effect of the body mass index is close to quadratic but the effect of the duration of diabetes would be better modeled as cubic.

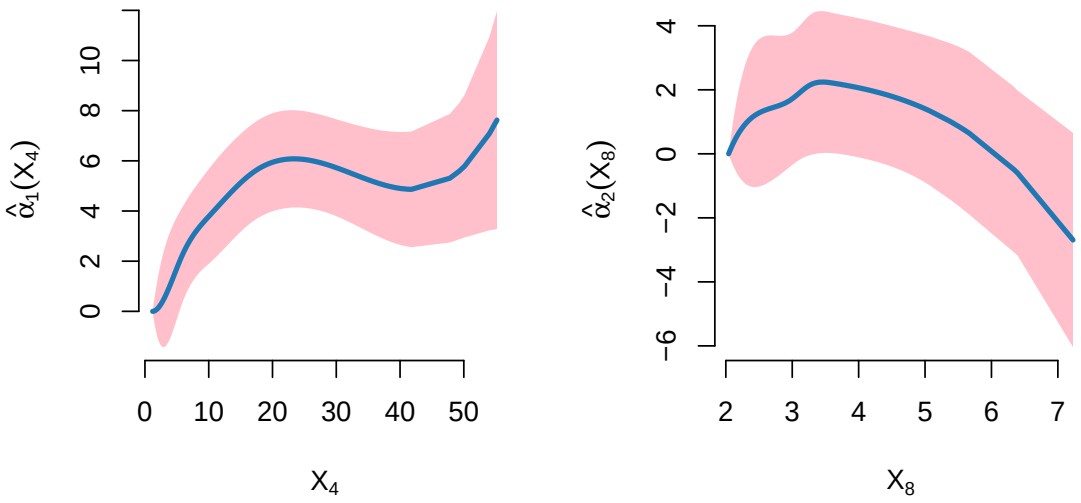

Figure 5.2: Analysis of the diabetic retinopathy data: estimated nonparametric functions and 95% pointwise confidence intervals.

6. Conclusions and Discussions

We have assumed that the number of covariates in the linear part diverges while that in the nonparametric part remains fixed. Our theory can be potentially generalized to the case when the latter also diverges. For example, the results stated in Theorem 3.1 are expected to hold if we add a multiplicative factor involving q to the definition of r_n, under conditions similar as (A1)–(A6).

With a large number of nonparametric components, it is desirable to perform variable selection for the nonparametric components as well as the parametric components. It is conceptually straightforward to extend our variable selection methodology using two penalties, resulting in a doubly penalized GEE:

$$\mathbf{L}(\mathbf{b}) = \mathbf{S}(\mathbf{b}) - n\mathbf{q}_{\lambda_1}(\|\mathbf{a}\|)\mathrm{sgn}(\mathbf{a}) - \mathrm{n}\mathbf{q}_{\lambda_2}(|\boldsymbol{\beta}|)\mathrm{sgn}(\boldsymbol{\beta}),$$

where $\mathrm{sgn}(\mathbf{a}) = (\mathbf{a}_1^{\mathrm{T}}/\|\mathbf{a}_1\|, \ldots, \mathbf{a}_{\mathrm{q}}^{\mathrm{T}}/\|\mathbf{a}_{\mathrm{q}}\|, \mathbf{0}_{\mathrm{p}}^{\mathrm{T}})^{\mathrm{T}}$ (by convention $\mathbf{a}_l/\|\mathbf{a}_l\|$ is $\mathbf{0}$ if $\mathbf{a}_l = \mathbf{0}$) and $\mathbf{q}_{\lambda_1}(\|\mathbf{a}\|) = (q_{\lambda_1}(\|\mathbf{a}_1\|), \ldots, q_{\lambda_1}(\|\mathbf{a}_q\|), \mathbf{0}_p^{\mathrm{T}})^{\mathrm{T}}$. As one needs to select λ_1 and λ_2 simultaneously, the computation is expected to be burdensome. Investigation of the computational and theoretical properties of this doubly penalized GEE will be left for future study.

In the dataset analysis, we have separated the covariates into the nonparametric part and parametric part in an initial screening stage based on visual inspection. Although this makes a sensible first attempt, a more efficient and automatic criterion is needed to determine which covariates should be included in the linear component and which in the nonparametric part. Recently Zhang et al. (2011) investigated this problem using two penalties where one penalty is designed to identify the linear part. It remains challenging to extend this work to longitudinal data with diverging p. This is a future research topic.

Automatic choice of the number of knots is an important issue in spline estimation. We have followed the advice in some previous studies, e.g. Huang et al. (2010), and fixed the number of knots. Empirically, we find that two or three internal knots are flexible enough to approximate smooth functions in most situations, although a larger number of knots is required for curves with more complicated shapes. In a high-dimensional setting, developing suitable criterion for knots selection is a challenging problem.

7. Proofs

Proof of Proposition 3.1 and a Preliminary Lemma.

(i) Let $\widetilde{\mathbf{H}} = \sum_i \mathbf{U}_i^{\mathrm{T}}\mathbf{V}_i\mathbf{U}_i$, where $\mathbf{V}_i = \mathbf{A}_{0i}^{1/2}\mathbf{R}_0\mathbf{A}_{0i}^{1/2}$ is the true covariance matrix.

Based on the injection lemma (Chen, Hu and Ying (1999)), we need only show that for any $\epsilon > 0$, there exists a $\Delta > 0$ such that, for n large enough,

$$P\left(\|\widetilde{\mathbf{H}}^{-1/2}\widetilde{\mathbf{S}}(\mathbf{b}_0)\| \leq \inf_{\|\mathbf{b}-\mathbf{b}_0\|=\Delta r_n} \|\widetilde{\mathbf{H}}^{-1/2}\{\widetilde{\mathbf{S}}(\mathbf{b}) - \widetilde{\mathbf{S}}(\mathbf{b}_0)\}\|\right) \geq 1-\epsilon. \tag{7.8}$$

Using a Taylor expansion, we have

$$\begin{aligned}
&\|\widetilde{\mathbf{H}}^{-1/2}\{\widetilde{\mathbf{S}}(\mathbf{b}) - \widetilde{\mathbf{S}}(\mathbf{b}_0)\}\| \\
&\qquad = \|\widetilde{\mathbf{H}}^{-1/2}[\sum_i \mathbf{U}_i^{\intercal}\{\boldsymbol{\mu}_i(\mathbf{b}) - \boldsymbol{\mu}_i(\mathbf{b}_0)\}]\| \\
&\qquad = \|\widetilde{\mathbf{H}}^{-1/2}\{\sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{U}_{ij}\dot{\mu}(\theta_{ij}^*)\mathbf{U}_{ij}^{\intercal}(\mathbf{b}-\mathbf{b}_0)\}\| \\
&\qquad \geq C\lambda_{\min}(\widetilde{\mathbf{H}}^{-1/2})\lambda_{\min}(\sum_i \mathbf{U}_i^{\intercal}\mathbf{U}_i)(\Delta r_n) \\
&\qquad \geq C\sqrt{n}\Delta r_n,
\end{aligned}$$

where $\mathbf{U}_{ij}$ is the j-th row of $\mathbf{U}_i$ and θ_{ij}^* lies between $\mathbf{U}_{ij}^{\intercal}\mathbf{b}$ and $\mathbf{U}_{ij}^{\intercal}\mathbf{b}_0$.

On the other hand,

$$\begin{aligned}
\|\widetilde{\mathbf{H}}^{-1/2}\widetilde{\mathbf{S}}(\mathbf{b}_0)\| &\leq \|\widetilde{\mathbf{H}}^{-1/2}\sum_i \mathbf{U}_i^{\intercal}(\mathbf{Y}_i - \boldsymbol{\mu}_{0i})\| + \|\widetilde{\mathbf{H}}^{-1/2}\sum_i \mathbf{U}_i^{\intercal}\{\boldsymbol{\mu}_{0i} - \boldsymbol{\mu}_i(\mathbf{b}_0)\}\| \\
&\leq \sqrt{\sum_i \boldsymbol{\epsilon}_i^{\intercal}\mathbf{U}_i\widetilde{\mathbf{H}}^{-1}\mathbf{U}_i^{\intercal}\boldsymbol{\epsilon}_i} + \|\widetilde{\mathbf{H}}^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{U}_{ij}\dot{\mu}(\theta_{ij}^*)(\mathbf{U}_{ij}^{\intercal}\mathbf{b}_0 - \theta_{0ij})\| \\
&= O_p(\sqrt{\mathrm{tr}(\mathbf{U}\widetilde{\mathbf{H}}^{-1}\mathbf{U}^{\intercal})}) + O_p(\|\mathbf{U}_{ij}^{\intercal}\mathbf{b}_0 - \theta_{0ij}\|)\sqrt{\lambda_{\max}(\mathbf{U}\widetilde{\mathbf{H}}^{-1}\mathbf{U}^{\intercal})} \\
&= O_p(\sqrt{K+p}) + O_p(\sqrt{n}K^{-d}) = O_p(\sqrt{n}r_n),
\end{aligned}$$

where θ_{ij}^* lies between $\mathbf{U}_{ij}^{\intercal}\mathbf{b}_0$ and θ_{0ij}, $\boldsymbol{\epsilon}_i = (\mathbf{Y}_i - \boldsymbol{\mu}_{0i})$, $\mathbf{U} = (\mathbf{U}_1^{\intercal}, \ldots, \mathbf{U}_n^{\intercal})^{\intercal}$. Thus (7.8) is proved if Δ is large enough.

(ii) The proof is basically the same as the proof for (3.4) in Wang (2011). Let

$$\mathbf{R}^* = \frac{1}{n}\sum_i \mathbf{A}_{0i}^{-1/2}(\mathbf{Y}_i - \boldsymbol{\mu}_{0i})(\mathbf{Y}_i - \boldsymbol{\mu}_{0i})^{\intercal}\mathbf{A}_{0i}^{-1/2}.$$

Using the Central Limit Theorem indicates $\|\mathbf{R}^* - \mathbf{R}_0\| = O_p(n^{-1/2})$. We can also show $\|\widehat{\mathbf{R}} - \mathbf{R}^*\| = O_p(r_n)$ following the proof of (3.4) in Wang (2011). Finally we have $\|\widehat{\mathbf{R}}^{-1} - \mathbf{R}_0^{-1}\| = \|\widehat{\mathbf{R}}^{-1}(\widehat{\mathbf{R}} - \mathbf{R}_0)\mathbf{R}_0^{-1}\| = O_p(r_n)$. □

Lemma 7.1. *If (A1)–(A6) hold,* $p/n \to 0, K\log K/n \to 0$ *and* $K \to \infty$*, then*

$$||\mathbf{S}(\mathbf{b}_0) - \overline{\mathbf{S}}(\mathbf{b}_0)|| = O_p(nr_n^2). \tag{7.9}$$

Furthermore, for $\mathbf{b} \in R^{qK+p}$*, we have*

$$\sup_{||\mathbf{b}-\mathbf{b}_0||\le\Delta r_n}\sup_{||\mathbf{d}||=1}\left|\mathbf{d}^{\mathrm{T}}\{\mathbf{D}(\mathbf{b}) - \overline{\mathbf{D}}(\mathbf{b})\}\mathbf{d}\right| = O_p(nr_n). \tag{7.10}$$

$$\sup_{||\mathbf{b}-\mathbf{b}_0||\le\Delta r_n}\sup_{||\mathbf{d}||=1}\left|\mathbf{d}^{\mathrm{T}}\{\overline{\mathbf{D}}(\mathbf{b}) - \overline{\mathbf{H}}(\mathbf{b})\}\mathbf{d}\right| = O_p\left(n(K+p)^{1/2}r_n\right). \tag{7.11}$$

$$\sup_{||\mathbf{b}-\mathbf{b}_0||\le\Delta r_n}\sup_{||\mathbf{d}||=1}\left|\mathbf{d}^{\mathrm{T}}\{\overline{\mathbf{H}}(\mathbf{b}) - \overline{\mathbf{H}}(\mathbf{b}_0)\}\mathbf{d}\right| = O_p\left(n(K+p)^{1/2}r_n\right). \tag{7.12}$$

Proof. Let $\mathbf{Q} = \{q_{j_1,j_2}\}_{1\le j_1,j_2\le m}$ denote the matrix $\widehat{\mathbf{R}}^{-1} - \overline{\mathbf{R}}^{-1}$. We can write

$$\mathbf{S}(\mathbf{b}_0) - \overline{\mathbf{S}}(\mathbf{b}_0) = \sum_{i=1}^{n}\sum_{j_1=1}^{m}\sum_{j_2=1}^{m} q_{j_1,j_2}\mathbf{A}_{ij_1}^{1/2}(\mathbf{b}_0)\mathbf{A}_{ij_2}^{-1/2}(\mathbf{b}_0)\{Y_{ij_2} - \mu_{ij_2}(\mathbf{b}_0)\}\mathbf{U}_{ij_1}.$$

Note that $Y_{ij_2} - \mu_{ij_2}(\mathbf{b}_0)$ does not have mean zero. We further decompose

$$\begin{aligned}\mathbf{S}(\mathbf{b}_0) - \overline{\mathbf{S}}(\mathbf{b}_0) &= \sum_{i=1}^{n}\sum_{j_1=1}^{m}\sum_{j_2=1}^{m} q_{j_1,j_2}\mathbf{A}_{ij_1}^{1/2}(\mathbf{b}_0)\mathbf{A}_{ij_2}^{-1/2}(\mathbf{b}_0)(Y_{ij_2} - \mu_{0ij_2})\mathbf{U}_{ij_1} \\ &\quad + \sum_{i=1}^{n}\sum_{j_1=1}^{m}\sum_{j_2=1}^{m} q_{j_1,j_2}\mathbf{A}_{ij_1}^{1/2}(\mathbf{b}_0)\mathbf{A}_{ij_2}^{-1/2}(\mathbf{b}_0)\{\mu_{0ij_2} - \mu_{ij_2}(\mathbf{b}_0)\}\mathbf{U}_{ij_1}.\end{aligned}$$

Then, as in Lemma 3.1 of Wang (2011), we have $\| \sum_{i=1}^{n} \mathbf{A}_{ij_1}^{1/2}(\mathbf{b}_0)\mathbf{A}_{ij_2}^{-1/2}(\mathbf{b}_0)(Y_{ij_2} - \mu_{0ij_2})\mathbf{U}_{ij_1}|| = O_p(\sqrt{n(K+p)})$. Furthermore,

$$\begin{aligned}&\| \sum_{i=1}^{n} \mathbf{A}_{ij_1}^{1/2}(\mathbf{b}_0)\mathbf{A}_{ij_2}^{-1/2}(\mathbf{b}_0)\{\mu_{0ij_2} - \mu_{ij_2}(\mathbf{b}_0)\}\mathbf{U}_{ij_1}|| \\ &\quad\le C\| \sum_{i=1}^{n}\{\mu_{0ij_2} - \mu_{ij_2}(\mathbf{b}_0)\}\mathbf{U}_{ij_1}|| \\ &\quad= O_p\left(\sqrt{\sum_i\{\mu_{0ij_2} - \mu_{ij_2}(\mathbf{b}_0)\}^2}\right) O_p\left(\sqrt{\lambda_{\max}(\sum_i \mathbf{U}_{ij_1}\mathbf{U}_{ij_2}^{\mathrm{T}})}\right) \\ &\quad= O_p(nK^{-d}).\end{aligned}$$

By Condition (A5), $q_{j_1,j_2} = O_p(r_n)$, $\forall\ 1 \le j_1, j_2 \le m$, and the proof of (7.9) is finished.

the proof of (7.10)-(7.12) can be done similarly largely as in Lemma C.2 of Wang (2011). □

Proof of Theorem 3.1.

As in the proof of Proposition 3.1, we need only show that for any $\epsilon > 0$, there exists a $\Delta > 0$ such that, for n large enough,

$$P\left(\|\overline{\mathbf{M}}^{-1/2}(\mathbf{b}_0)\mathbf{S}(\mathbf{b}_0)\| \leq \inf_{\|\mathbf{b}-\mathbf{b}_0\|=\Delta r_n} \|\overline{\mathbf{M}}^{-1/2}(\mathbf{b}_0)\{\mathbf{S}(\mathbf{b}) - \mathbf{S}(\mathbf{b}_0)\}\|\right) \geq 1-\epsilon. \tag{7.13}$$

Considering first the right hand side of the event, we have

$$\begin{aligned}
\overline{\mathbf{M}}^{-1/2}(\mathbf{b}_0)\{\mathbf{S}(\mathbf{b}) - \mathbf{S}(\mathbf{b}_0)\} &= \overline{\mathbf{M}}^{-1/2}(\mathbf{b}_0)\mathbf{D}(\mathbf{b}^*)(\mathbf{b}-\mathbf{b}_0) \\
&= \overline{\mathbf{M}}^{-1/2}(\mathbf{b}_0)\overline{\mathbf{H}}(\mathbf{b}_0)(\mathbf{b}-\mathbf{b}_0) + \overline{\mathbf{M}}^{-1/2}(\mathbf{b}_0)\{\overline{\mathbf{H}}(\mathbf{b}^*) - \overline{\mathbf{H}}(\mathbf{b}_0)\}(\mathbf{b}-\mathbf{b}_0) \\
&\quad + \overline{\mathbf{M}}^{-1/2}(\mathbf{b}_0)\{\overline{\mathbf{D}}(\mathbf{b}^*) - \overline{\mathbf{H}}(\mathbf{b}^*)\}(\mathbf{b}-\mathbf{b}_0) \\
&\quad + \overline{\mathbf{M}}^{-1/2}(\mathbf{b}_0)\{\mathbf{D}(\mathbf{b}^*) - \overline{\mathbf{D}}(\mathbf{b}^*)\}(\mathbf{b}-\mathbf{b}_0) \\
&\triangleq \sum_{j=1}^{4} I_{nj},
\end{aligned}$$

where $\mathbf{b}^*$ lies between $\mathbf{b}$ and $\mathbf{b}_0$. We have $\|I_{n1}\| \geq C\sqrt{n}\|\mathbf{b}-\mathbf{b}_0\| = C\sqrt{n}\Delta r_n$, and

$$\begin{aligned}
\|I_{n2}\| &= \|\mathbf{b}-\mathbf{b}_0\| \max[|\lambda_{\max}\{\overline{\mathbf{H}}(\mathbf{b}^*) - \overline{\mathbf{H}}(\mathbf{b}_0)\}|, |\lambda_{\min}\{\overline{\mathbf{H}}(\mathbf{b}^*) - \overline{\mathbf{H}}(\mathbf{b}_0)\}|]/\lambda^{1/2}_{\max}\{\overline{\mathbf{M}}(\mathbf{b}_0)\} \\
&= O_p(r_n)O_p(nr_n\sqrt{K+p})O_p(1/\sqrt{n}) \\
&= o_p(\sqrt{n}r_n).
\end{aligned}$$

Similarly we get $\|I_{nj}\| = o_p(\sqrt{n}r_n), j = 3, 4$.

For the left hand side of the event in (7.13),

$$\begin{aligned}
\|\overline{\mathbf{M}}^{-1/2}(\mathbf{b}_0)\mathbf{S}(\mathbf{b}_0)\| &\leq \|\overline{\mathbf{M}}^{-1/2}(\mathbf{b}_0)\{\mathbf{S}(\mathbf{b}_0) - \overline{\mathbf{S}}(\mathbf{b}_0)\}\| \\
&\quad + \|\overline{\mathbf{M}}^{-1/2}(\mathbf{b}_0)\sum_i \mathbf{U}_i^{\tau}\mathbf{A}_i^{1/2}(\mathbf{b}_0)\overline{\mathbf{R}}^{-1}\mathbf{A}_i^{-1/2}(\mathbf{b}_0)\{\boldsymbol{\mu}_{0i} - \boldsymbol{\mu}_i(\mathbf{b}_0)\}\| \\
&\quad + \|\overline{\mathbf{M}}^{-1/2}(\mathbf{b}_0)\sum_i \mathbf{U}_i^{\tau}\mathbf{A}_i^{1/2}(\mathbf{b}_0)\overline{\mathbf{R}}^{-1}\mathbf{A}_i^{-1/2}(\mathbf{b}_0)(\mathbf{Y}_i - \boldsymbol{\mu}_{0i})\| \\
&\triangleq J_{n1} + J_{n2} + J_{n3}.
\end{aligned}$$

We have $\|J_{n1}\| = o_p(\sqrt{n}r_n)$ by Lemma 7.1, and

$$\|J_{n2}\| = O_p(1/\sqrt{n})O_p(\sqrt{n}K^{-d})O_p\left(\sqrt{\lambda_{\max}(\sum_i \mathbf{U}_i\mathbf{U}_i^{\mathrm{T}})}\right) = O_p(\sqrt{n}K^{-d}) = O_p(\sqrt{n}r_n).$$

By straightforward calculations $EJ_{n3} = 0$ and $E\|J_{n3}\|^2 = K + p$, so $\|J_{n3}\| = O_p(\sqrt{K+p}) = O_p(\sqrt{n}r_n)$.

Combining these results, (7.13) is proved with Δ large enough. □

Proof of Theorem 3.2.

Since the entries of Γ, denoted by $h_{jl}^{(s)}$, are in $\mathfrak{H}_d$, $h_{jl}^{(s)}$ can be approximated by spline functions $\tilde{h}_{jl}^{(s)}$ with approximation error $O(K^{-d})$. Denote by $\widehat{\Gamma}(\mathbf{W}_i)$ the matrix that approximates $\Gamma(\mathbf{W}_i)$ by replacing $h_{jl}^{(s)}$ with $\tilde{h}_{jl}^{(s)}$. Note that since $\tilde{h}_{jl}^{(s)}$ is a spline function, the (j,s)-entry of $\widehat{\Gamma}(\mathbf{W}_i)$ can be expressed as $\mathbf{c}_j^{(s)}\mathbf{B}_{ij}$ for some $\mathbf{c}_j^{(s)} \in R^{qK}$, for $1 \le j \le m, 1 \le s \le p$.

Let $\widehat{\mathbf{b}}_0 = (\hat{\mathbf{a}}^{\mathrm{T}}, \boldsymbol{\beta}_0^{\mathrm{T}})^{\mathrm{T}}$ and $\hat{\mathbf{b}} = (\hat{\mathbf{a}}^{\mathrm{T}}, \hat{\boldsymbol{\beta}}^{\mathrm{T}})^{\mathrm{T}}$. Take

$$\begin{aligned}
\mathcal{S}(\mathbf{b}) &= \sum_i \{\mathbf{X}_i - \widehat{\Gamma}(\mathbf{W}_i)\}^{\mathrm{T}}\mathbf{A}_i^{1/2}(\mathbf{b})\widehat{\mathbf{R}}^{-1}\mathbf{A}_i^{-1/2}(\mathbf{b})\{\mathbf{Y}_i - \boldsymbol{\mu}_i(\mathbf{b})\},\\
\overline{\mathcal{S}}(\mathbf{b}) &= \sum_i \{\mathbf{X}_i - \widehat{\Gamma}(\mathbf{W}_i)\}\mathbf{A}_i^{1/2}(\mathbf{b})\overline{\mathbf{R}}^{-1}\mathbf{A}_i^{-1/2}(\mathbf{b}_0)\{\mathbf{Y}_i - \boldsymbol{\mu}_i(\mathbf{b})\},\\
\overline{\mathcal{S}}_0 &= \sum_i \{\mathbf{X}_i - \Gamma(\mathbf{W}_i)\}^{\mathrm{T}}\mathbf{A}_{0i}^{1/2}\overline{\mathbf{R}}^{-1}\mathbf{A}_{0i}^{-1/2}(\mathbf{Y}_i - \boldsymbol{\mu}_{0i}),
\end{aligned}$$

$\mathcal{D}(\mathbf{b}) = -\frac{\partial\mathcal{S}(\mathbf{a},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{\mathrm{T}}}$, $\overline{\mathcal{D}}(\mathbf{b}) = -\frac{\partial\overline{\mathcal{S}}(\mathbf{a},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{\mathrm{T}}}$, and

$$\overline{\mathcal{H}}(\mathbf{b}) = \{\mathbf{X}_i - \widehat{\Gamma}(\mathbf{W}_i)\}^{\mathrm{T}}\mathbf{A}_i^{1/2}(\mathbf{b})\overline{\mathbf{R}}^{-1}\mathbf{A}_i^{1/2}(\mathbf{b})\{\mathbf{X}_i - \widehat{\Gamma}(\mathbf{W}_i)\}.$$

We need $\mathcal{S}(\widehat{\mathbf{b}}) = 0$. To see this, note that

$$\mathbf{S}(\widehat{\mathbf{b}}) = \sum_i \mathbf{U}_i^{\mathrm{T}}\mathbf{A}_i^{1/2}(\widehat{\mathbf{b}})\widehat{\mathbf{R}}^{-1}\mathbf{A}_i^{1/2}(\widehat{\mathbf{b}})\{\mathbf{Y}_i - \boldsymbol{\mu}_i(\widehat{\mathbf{b}})\} = 0,$$

and the (j,s)-entry of $\widehat{\Gamma}(\mathbf{W}_i)$ can be expressed as $\mathbf{c}_j^{(s)}\mathbf{B}_{ij}$. Thus it is easy to see that the rows of $\mathcal{S}(\widehat{\mathbf{b}})$ are simply linear combinations of the rows of $\mathbf{S}(\widehat{\mathbf{b}})$, which implies $\mathcal{S}(\widehat{\mathbf{b}}) = 0$.

By the arguments in Lemma 3.7 in Wang (2011), $\boldsymbol{\alpha}^{\mathrm{T}}\overline{\mathcal{M}}_0^{-1/2}\overline{\mathcal{S}}_0 \to N(0,1)$. We next show that

$$\|\mathcal{S}(\widehat{\mathbf{b}}_0) - \overline{\mathcal{S}}_0\| = o_p(\sqrt{n}). \tag{7.14}$$

In fact,

$$\begin{aligned}\|\mathcal{S}(\widehat{\mathbf{b}}_0)-\overline{\mathcal{S}}_0\| \le & \|\sum_i\{\mathbf{X}_i-\widehat{\Gamma}(\mathbf{W}_i)\}\mathbf{A}_i^{1/2}(\widehat{\mathbf{b}}_0)\widehat{\mathbf{R}}^{-1}\mathbf{A}_i^{-1/2}(\widehat{\mathbf{b}}_0)\{\boldsymbol{\mu}_i(\widehat{\mathbf{b}}_0)-\boldsymbol{\mu}_{0i}\}\| \\ & +\|\sum_i\{\mathbf{X}_i-\widehat{\Gamma}(\mathbf{W}_i)\}\mathbf{A}_i^{1/2}(\widehat{\mathbf{b}}_0)(\widehat{\mathbf{R}}^{-1}-\overline{\mathbf{R}}^{-1})\mathbf{A}_i^{-1/2}(\widehat{\mathbf{b}}_0)(\mathbf{Y}_i-\boldsymbol{\mu}_{0i})\| \\ & +\|\sum_i\{\widehat{\Gamma}(\mathbf{W}_i)-\Gamma(\mathbf{W}_i)\}\mathbf{A}_i^{1/2}(\widehat{\mathbf{b}}_0)\overline{\mathbf{R}}^{-1}\mathbf{A}_i^{-1/2}(\widehat{\mathbf{b}}_0)(\mathbf{Y}_i-\boldsymbol{\mu}_{0i})\| \\ & +\|\sum_i\{\mathbf{X}_i-\Gamma(\mathbf{W}_i)\}(\mathbf{A}_i^{1/2}(\widehat{\mathbf{b}}_0)-\mathbf{A}_{0i}^{1/2})\overline{\mathbf{R}}^{-1}\mathbf{A}_i^{-1/2}(\widehat{\mathbf{b}}_0)(\mathbf{Y}_i-\boldsymbol{\mu}_{0i})\| \\ & +\|\sum_i\{\mathbf{X}_i-\Gamma(\mathbf{W}_i)\}\mathbf{A}_{0i}^{1/2}\overline{\mathbf{R}}^{-1}\{\mathbf{A}_i^{-1/2}(\widehat{\mathbf{b}}_0)-\mathbf{A}_{0i}^{-1/2}\}(\mathbf{Y}_i-\boldsymbol{\mu}_{0i})\| \\ \stackrel{\triangle}{=} & \sum_{j=1}^5 L_{nj}.\end{aligned}$$

First,

$$\begin{aligned}L_{n1} \le & \|\sum_i\{\mathbf{X}_i-\Gamma(\mathbf{W}_i)\}\mathbf{A}_i^{1/2}(\widehat{\mathbf{b}}_0)\widehat{\mathbf{R}}^{-1}\mathbf{A}_i^{-1/2}(\widehat{\mathbf{b}}_0)\{\boldsymbol{\mu}_i(\widehat{\mathbf{b}}_0)-\boldsymbol{\mu}_{0i}\}\| \\ & +\|\sum_i\{\Gamma(\mathbf{W}_i)-\widehat{\Gamma}(\mathbf{W}_i)\}\mathbf{A}_i^{1/2}(\widehat{\mathbf{b}}_0)\widehat{\mathbf{R}}^{-1}\mathbf{A}_i^{-1/2}(\widehat{\mathbf{b}}_0)\{\boldsymbol{\mu}_i(\widehat{\mathbf{b}}_0)-\boldsymbol{\mu}_{0i}\}\|.\end{aligned}$$

From the definition of $\Gamma(\mathbf{W}_i)$, as the proof of (A.6) in Wang et al. (2011), the first term above is $O_p(\sqrt{np}r_n)$. The second term is obviously $O_p(n\sqrt{p}K^{-d}r_n)$ since $\|\Gamma(\mathbf{W}_i)-\widehat{\Gamma}(\mathbf{W}_i)\|=O_p(\sqrt{p}K^{-d})$ and $\|\boldsymbol{\mu}_i(\widehat{\mathbf{b}}_0)-\boldsymbol{\mu}_{0i}\|=O_p(r_n)$. Thus if $\frac{np}{K^{2d}}(\frac{K+p}{n}+\frac{1}{K^{2d}})=o(1)$, then $L_{n1}=o_p(\sqrt{n})$. Using Condition (A1) and Proposition 3.1, we get $L_{n2}=o_p(\sqrt{n})$. Similarly one can get $L_{nj}=o_p(\sqrt{n})$, $j=3,4,5$, and thus (7.14) is shown.

Based on (7.14) and using a Taylor expansion, we get

$$\begin{aligned}\boldsymbol{\alpha}^{\mathrm{T}}\overline{\mathcal{M}}_{0n}^{-1/2}\overline{\mathcal{S}}_0 &= \boldsymbol{\alpha}^{\mathrm{T}}\overline{\mathcal{M}}_{0n}^{-1/2}\mathcal{S}(\widehat{\mathbf{b}}_0)+o_p(1) \\ &= \boldsymbol{\alpha}^{\mathrm{T}}\overline{\mathcal{M}}_{0n}^{-1/2}\mathcal{D}(\mathbf{b}^*)(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)+o_p(1), \qquad (7.15)\end{aligned}$$

where $\mathbf{b}^*=(\widehat{\mathbf{a}},\boldsymbol{\beta}^*)$ with $\boldsymbol{\beta}^*$ lying between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$, and in the last step above we used $\mathcal{S}(\widehat{\mathbf{b}})=0$.

Using the arguments in the proof of Lemma 7.1 (or of Lemma C.2 in Wang et al. (2011)), we can show that

$$\sup_{||\mathbf{b}-\mathbf{b}_0||\le\Delta r_n}\sup_{||\mathbf{d}||=1}\left|\mathbf{d}^{\mathrm{T}}\{\mathcal{D}(\mathbf{b})-\overline{\mathcal{D}}(\mathbf{b})\}\mathbf{d}\right| = O_p(nr_n).$$

$$\sup_{||\mathbf{b}-\mathbf{b}_0||\le\Delta r_n}\sup_{||\mathbf{d}||=1}\left|\mathbf{d}^{\tau}\{\overline{\mathcal{D}}(\mathbf{b})-\overline{\mathcal{H}}(\mathbf{b})\}\mathbf{d}\right| = O_p\left(np^{1/2}r_n\right).$$
$$\sup_{||\mathbf{b}-\mathbf{b}_0||\le\Delta r_n}\sup_{||\mathbf{d}||=1}\left|\mathbf{d}^{\tau}\{\overline{\mathcal{H}}(\mathbf{b})-\overline{\mathcal{H}}(\mathbf{b}_0)\}\mathbf{d}\right| = O_p\left(np^{1/2}r_n\right).$$

Thus (7.15) comes to

$$\boldsymbol{\alpha}^{\tau}\overline{\mathcal{M}}_0^{-1/2}\overline{\mathcal{S}}_0=\boldsymbol{\alpha}^{\tau}\overline{\mathcal{M}}_0^{-1/2}\overline{\mathcal{H}}(\mathbf{b}_0)(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)+o_p(1),$$

from which the asymptotic normality follows by the assumptions imposed on K.

Proof of Theorem 4.3.

Let $(\hat{\mathbf{a}}^{o\tau},\hat{\boldsymbol{\beta}}^{(1)\tau})^{\tau}$ be the exact solution of (2.3) when we only use the first s columns of $\mathbf{X}_i$ in the definition $\mathbf{S}_i$. We show that $\hat{\mathbf{b}}=(\hat{\mathbf{a}}^{o\tau},\hat{\boldsymbol{\beta}}^{(1)\tau},\hat{\boldsymbol{\beta}}^{(2)\tau}=\mathbf{0})^{\tau}$ satisfies all the properties in the theorem.

Property (i) is trivial by our construction. (ii) is also obvious as $(\hat{\mathbf{a}}^{o},\hat{\boldsymbol{\beta}}^{(1)})$ solves (2.3) and the penalty term is zero since $\min_{1\le j\le s}|\hat{\beta}_j|/\lambda\to\infty$. Based on Theorems 3.1 and 3.2, the convergence rate is $O(r_n)$ and $\hat{\boldsymbol{\beta}}^{(1)}$ is asymptotically normal. Thus we only need to demonstrate (iii).

Motivated by the proof of Theorem 3.2, we let

$$\begin{aligned}&\mathcal{S}_{\hat{\mathbf{a}}^o}(\boldsymbol{\beta})\\ =\;&\sum_i\{\mathbf{X}_i-\widehat{\Gamma}(\mathbf{W}_i)\}^{\tau}\mathbf{A}_i^{1/2}((\hat{a}^{o\tau},\boldsymbol{\beta}^{\tau})^{\tau})\widehat{\mathbf{R}}^{-1}\mathbf{A}_i^{-1/2}((\hat{a}^{o\tau},\boldsymbol{\beta}^{\tau})^{\tau})\{\mathbf{Y}_i-\boldsymbol{\mu}_i((\hat{a}^{o\tau},\boldsymbol{\beta}^{\tau})^{\tau})\}.\end{aligned}$$

Similar to (7.14) and using the root-n consistency of $\hat{\boldsymbol{\beta}}^{(1)}$, we have $\|\mathcal{S}_{\hat{\mathbf{a}}^o}((\hat{\boldsymbol{\beta}}^{(1)\tau},\hat{\boldsymbol{\beta}}^{(2)\tau})^{\tau})-\bar{\mathcal{S}}_0\|=o_p(\sqrt{n})=o_p(n\lambda)$. Using the arguments in the proof of Theorem 1 in Wang et al. (2012), with the only difference being that $\mathbf{X}_i$ in their paper is replaced by $\mathbf{X}_i-\Gamma(\mathbf{W}_i)$ here, we have $\sup_{s+1\le j\le p}\bar{\mathcal{S}}_{0j}=O_p(n\lambda/\log n)$, where $\bar{\mathcal{S}}_{0j}$ is the jth component of $\bar{\mathcal{S}}_0$. Thus uniformly over $j=pK+1,\ldots,pK+s$, $\mathbf{S}_j(\hat{\mathbf{b}}+\epsilon\mathbf{e}_j)$ and $\mathbf{S}_j(\hat{\mathbf{b}}-\epsilon\mathbf{e}_j)$ are dominated by $-nq_\lambda(\epsilon)$ and $nq_\lambda(\epsilon)$ respectively, as ϵ goes to zero. Thus $\mathbf{L}_j(\hat{\mathbf{b}}+\epsilon\mathbf{e}_j)$ and $\mathbf{L}_j(\hat{\mathbf{b}}-\epsilon\mathbf{e}_j)$ have different signs. This completes the proof of (iii). □

Acknowledgments

The research was partially supported by Singapore MOE Tier 1 RG 62/11 (Lian); NSF grant DMS-1007167 (Liang), and NSF grant DMS-1007603 (Wang). We thank the

Editor, an AE and two referees for their helpful comments which helped us to significantly improve the paper.

References

Barnhart, H. X. and Williamson, J. M. (1998). Goodness-of-fit tests for GEE modeling with binary responses. *Biometrics* **54**, 720-729.

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion). *The Annals of Statistics* **17**, 453-555.

Cantoni, E., Flemming, J. M. and Ronchetti, E. (2005). Variable selection for marginal longitudinal generalized linear models. *Biometrics* **61**, 507-514.

Chen, J. and Chen, Z. (2008). Extended bayesian information criterion for model selection with large model space. *Biometrica* **95**, 759-771.

Chen, K., Hu, I. and Ying, Z. (1999). Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *The Annals of Statistics* **27**, 1155-1163.

De Boor, C. (2001). *A Practical Guide to Splines* (Rev. ed.). New York: Springer-Verlag.

Diggle, P., Heagerty, P., Liang, K. and Zeger, S. (2002). *Analysis of Longitudinal Data* (2 ed.). Oxford: Oxford University Press.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* **20**, 101-148.

Fan, J. Q. and Li, R.(2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348-1360.

Fan, J. Q. and Li, R.(2006). Statistical challenges with high dimensionality: feature selection in knowledge discovery. In *International Congress of Mathematicians. Vol. III*, pp. 595-622. Eur. Math. Soc., Zurich.

Fitzmaurice, G. M. (1995). A caveat concerning independence estimating equations with multivariate binary data. *Biometrics* **51**, 309-317.

Hardle, W., Huet, S., Mammen, E. and Sperlich, S. (2004). Bootstrap inference in semiparametric generalized additive models. *Econometric Theory* **20**, 265-300.

Hardle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004). *Nonparametric and Semiparametric Models*. New York: Springer-Verlag.

He, X., Fung, W. and Zhu, Z. (2005). Robust estimation in generalized partial linear models for clustered data. *Journal of the American Statistical Association* **100**, 1176-1184.

Huang, J., Horowitz, J. L. and Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of Statistics* **38**, 2282-2313.

Hunter, D. R. and Li, R. (2005). Variable selection using mm algorithms. *Annals of Statistic* **33**, 1617-1642.

Johnson, B., Lin, D. and Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association* **103**, 672-680.

Klein, R., Klein, B. E. K., Moss, S. E., Davis, M. D. and DeMets, D. L. (1984). The Wisconsin Epidemiologic Study of Diabetic Retinopathy: III. Prevalence and Risk of Diabetic Retinopathy When Age at Diagnosis Is 30 or More Years. *Arch Ophthalmol* **102**, 527-532.

Li, Q. (2000). Efficient estimation of additive partially linear models. *International Economic Review* **41**, 1073-1092.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

Lin, X. and Carroll, R. J.(2001). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association* **96**, 1045-1056.

Lin, X. and Carroll, R. J. (2006). Semiparametric estimation in general repeated measures problems. *Journal of the Royal Statistical Society, Series B* **68**, 69-88.

Linton, O. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82**, 93-101.

Ma, S., Song, Q. and Wang, L. (2012). Simultaneous variable selection and estimation in semiparametric modeling of longitudinal/clustered data. *Bernoulli* to appear.

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *The Annals of Statistics* **14**, 590-606.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *The Annals of Statistics* **22**, 118-184.

Wang, L. (2011). GEE analysis of clustered binary data with diverging number of covariates. *The Annals of Statistics* **39**, 389-417.

Wang, L., X. Liu., H. Liang, and Carroll, R. J.(2011). Estimation and variable selection for generalized additive partial linear models. *The Annals of Statistics* **39**, 1827-1851.

Wang, L. and Qu, A.(2009). Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach. *Journal of the Royal Statistical Society, Series B* **71**, 177-190.

Wang, L., Zhou, J. and Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics* **68**, 353-360.

Wang, N. (2003). Marginal nonparametric kernel regression accounting for within subject correlation. *Biometrika* **90**, 43-52.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association* **99**, 673-686.

Wood, S. N. (2006). *Generalized Additive Models*. Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL.

Xue, L., Qu, A. and Zhou, J. (2010). Consistent model selection for marginal generalized additive model for correlated data. *Journal of the American Statistical Association* **105**, 1518-1530.

Yu, K., Park, B. U. and Mammen, E. (2008). Smooth backfitting in generalized additive models. *The Annals of Statistics* **36**, 228-260.

Zhang, H., Cheng, G. and Liu, Y. (2011). Linear or nonlinear? automatic structure discovery for partially linear models. *Journal of the American Statistical Association* **106**, 1099-1112.

Zhu, Z., Fung, W. K. and He, X. (2008). On the asymptotics of marginal regression splines with longitudinal data. *Biometrika* **95**, 907-917.

Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Singapore.

E-mail: (henglian@ntu.edu.sg)

Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, New York 14642, U.S.A.

E-mail: (hliang@bst.rochester.edu)

School of Statistics, University of Minnesota, Minneapolis, MN 55455, U.S.A.

E-mail: (wangx346@umn.edu)